



## REVIEW ARTICLE

Mayla Daiane Correa Molinari<sup>1\*</sup>   
Renata Fuganti-Pagliarini<sup>1</sup>   
Jéssika Angelotti Mendonça<sup>1</sup>   
Daniel de Amorim Barbosa<sup>1</sup>   
Daniel Rockenbach Marin<sup>1</sup>   
Liliane Mertz-Henning<sup>1</sup>   
Alexandre Lima Nepomuceno<sup>1</sup> 

<sup>1</sup> Embrapa Soybean, Carlos João Strass Road, Orlando Amaral Access, 86001-970, Londrina, Paraná, Brazil.

\* **Autora correspondente:**  
E-mail: maylamolinari@hotmail.com

### KEYWORDS

Sequencing  
Bioinformatics  
Pipeline  
RNA

### PALAVRAS-CHAVE

Sequenciamento  
Bioinformática  
Pipeline  
RNA

## Transcriptome analysis using RNA-Seq from experiments with and without biological replicates: a review

### *Análise de transcriptoma de experimentos de RNA-Seq com e sem repetições biológicas: revisão*

**ABSTRACT:** The discovery of nucleic acids opened new frontiers of knowledge, enabling researchers to access an enormous amount of data, through large-scale sequencing methodologies and bioinformatics tools. Amongst these new possibilities, RNA-Seq has been used to identify and quantify RNA molecules. To obtain more accurate biological responses from RNA-Seq data some questions should be considered such as experimental design, type of synthesized library, size of the fragments generated, number of biological replicates, depth, and coverage of the sequencing, species genome availability and, the choice of software to properly perform the computational analyzes. Accurate bioinformatics analyzes allow the selection of genes with a lower error rate, increasing the validation assertiveness via RT-qPCR and thus, reducing costs. The objective of this review was to present the analysis stages of RNA-Seq data, from experimental design to systems biology, considering relevant points, as well as to pointed out some software currently available to carry these analyzes out. Besides, with this review, we aimed to help the academic community to understand all steps and biases involved in RNA-Seq data analysis, from experiments with or without biological replicates.

**RESUMO:** A descoberta de ácidos nucleicos abriu novas fronteiras de conhecimento, permitindo que os pesquisadores acessassem uma enorme quantidade de dados, através de metodologias de sequenciamento em larga escala e ferramentas de bioinformática. Entre essas novas possibilidades, o RNA-Seq (sequenciamento de RNA) tem sido usado para identificar e quantificar moléculas de RNA. Para obter respostas biológicas mais precisas a partir dos dados de RNA-Seq, algumas questões devem ser consideradas, como o desenho experimental, o tipo de biblioteca sintetizada, o tamanho dos fragmentos gerados, o número de repetições biológicas, a profundidade e cobertura do sequenciamento, a disponibilidade do genoma da espécie e, a escolha dos softwares para executar adequadamente as análises computacionais. Análises bioinformáticas precisas permitem a seleção de genes com menor taxa de erro, aumentando a assertividade da validação via RT-qPCR e, assim, reduzindo custos. O objetivo desta revisão foi apresentar as etapas de análise de dados de RNA-Seq, desde o projeto experimental até a biologia dos sistemas, considerando pontos relevantes, bem como apontar alguns softwares atualmente disponíveis para realizar essas análises. Além disso, com esta revisão, objetivamos ajudar a comunidade acadêmica a compreender todas as etapas e vieses envolvidos na análise de dados de RNA-Seq, a partir de experimentos com ou sem réplicas biológicas.

Received: 13/06/2020  
Accepted: 29/12/2020

# 1 Introduction

Since 1953, when the structure of DNA was unveiled, this molecule has become one of the most studied in the world. At the beginning of the 70s/80s, DNA sequencing was time-consuming and labor-intensive, however, even with some restrictions, it opened a broad of new frontiers of knowledge (Wang *et al.*, 2009). Since 2005, nevertheless, new DNA and RNA sequencing tools were implemented and denominated Next Generation Sequencing (NGS), allowing a new large-scale sequencing approach (HTS-High Throughput Sequencing) (Heather & Chain, 2016). These new methodologies have considerably increased the scale and resolution of the analyzes. Besides, the reduction of the amount of sample needed and the significant reduction in nucleotide sequencing costs allowed the use of NGS directly in the sequencing of complete genomes, metagenomes, RNA-seq, exomes, non-coding RNAs, immunoprecipitations (ChIP-Seq), among several other applications (Wang *et al.*, 2009; Muhammad *et al.*, 2020).

To keep up with the progress of sequencing techniques and the new necessities to analyze the enormous amount of data generated, new science has emerged, named bioinformatics, an interdisciplinary field that corresponds to the application of computer techniques applied to biology (Quail *et al.*, 2012). To analyze a large amount of genetic sequencing data, bioinformatics was introduced in Brazil in 1999, through the complete DNA sequencing of the bacteria *Xylella fastidiosa*, a pathogen that causes serious damage to citrus crops (Simpson *et al.*, 2000). Since then, new software and analyzes tools have emerged.

Currently, the available bioinformatics tools allow researchers to analyze sets of genetic data from different cells, organs, and treatments quickly. At the end of the process allows the selection of interesting genes for validation but the most important is to select the correct pipeline to decrease false positives and let the analyzes in laboratories be rationalized to save money and time in lab techniques. This is possible based on an experimental design that represents the biological process of interest and enables better use of data, and robust data analyzes using a bioinformatics pipeline that uses the most appropriate software according to design. Both are necessary to influence obtaining true biological responses and, finally, a high correlation between the results of RNA-Seq and RT-qPCR.

To improve the correlation results between RNA-Seq and RT-qPCR, different software were tested for different experimental designs (Camarena *et al.*, 2010; Feng *et al.*, 2010; Zhang *et al.*, 2018; Wang *et al.*, 2019). Everaert *et al.* (2017) compared RNA-sequencing reads processed using five different workflows (Tophat-HTSeq, Tophat-Cufflinks, STAR-HTSeq, Kallisto, and Salmon) with gene expression levels generated by qPCR assays showing different accuracies according to experiment. These reports strengthen the robustness of RNA-Seq as a trustworthy methodology for transcripts

analyzes and gene expression quantification and reinforce the importance of the use of the bioinformatics tools more appropriate to experimental design (Conesa *et al.*, 2016).

In this review, to help ta academic community to understand the steps necessary to analyze RNA-Seq data according to experimental design all steps necessary were covered. From sequencing platform to biology system analysis, the authors described all stages of a pipeline and the main concerns that must be considered in each of the steps to perform robust and reliable analyzes in experiments with or without biological replicates. Some bioinformatics tools for data refinement have also been mentioned to show the numerous possibilities for analyzes that can be carried out under systems biology.

## 2 Material and Methods

### 2.1 Transcriptome analyzes through RNA-Seq

There are several techniques to perform transcriptional analyzes, some are based on hybridization (Microarrays and EST - expressed sequence tag) and others based on RNA sequencing (RNA-Seq). A more detailed comparison, covering aspects such as resolution, noise, amount of sample required, and cost, among these different methodologies used in the analyzes of the transcriptome can be seen in Table 1.

**Table 1.** Comparison between RNA-Seq, microarray, and EST technologies (Wang *et al.*, 2009, adapted by the authors).

**Tabela 1.** Comparação entre as tecnologias de RNA-Seq, microarranjo e EST (Wang *et al.*, 2009, adaptado pelos autores).

Specifications	Technology		
	Microarray	EST	RNA-Seq
Principle	Hybridization	Sanger	High throughput
Resolution	>100 bp*	Single base	Single base
Data quantity generated	High	Low	High
Use of a reference	Yes	No	No
Bias	High	Low	Low
Sample amount	High	High	Low
Cost	High	High	Low

\*bp = base pairs.

Nowadays, by comparison with other techniques, the RNA-Seq is the best cost/benefice tool available to perform transcriptome analyzes (Oshlack *et al.*, 2010). The advantages of RNA-Seq techniques are directed related to high-throughput sequencing technology, which allows researchers to sequence RNA on a large scale, faster, and with lower cost compared to other transcriptome analyzes techniques (Conesa *et al.*, 2016).

The RNA-Seq methodology also presents some gains up to other techniques.: 1. RNA-Seq does not depend on prior knowledge of the reference genome from the studied organism; 2. Reveals the precise localization of transcription limits; 3. Shows sequence variation within one base precision; 4. It is sensible to detect several expression levels, and 5. RNA-Seq does not need high RNA concentration to perform sequencing (Wang *et al.*,

2009).

In the last 10 years, the use of RNA-Seq as a technique of gene expression analyzes has increased significantly. One of its applications is to screen gene expression, in a specific development moment or an answer to a specific stimulus whether biotic or abiotic conditions. Reports of RNA-Seq use in transcriptome analyzes are available in a wide range of living organisms. In plants, examples of RNA-Seq being used to assay transcriptome analyzes can be found for *Arabidopsis thaliana*, *Zygophyllum xanthoxylum*, a xerophyte succulent, maize, rice and wheat, faba bean, oat, raspberry, sweet potato, barley, and soybean (Marcolino *et al.*, 2014; Nakayama *et al.*, 2017; Sharmin *et al.*, 2020).

Figure 1 below shows an overview of an RNA-Seq experiment and data analysis. All stages are illustrated, from the experiment itself to RNA extraction, the sequencing, and both bioinformatics analyses possibilities: *De novo* assembly (when a reference genome is not available and/or when the objective is to discover new transcripts), and Reference assembly (when reads are aligned in a known reference genome from your specie of interest).

## 2.2 Sequencing platforms for RNA-Seq experiments

The screening of differentially expressed genes requires the use of robust computational tools to allow the improvement of search to candidate genes on transcriptome (Prosdociami *et al.*, 2012). In short, through the transcriptome analysis is possible to (1) catalog all types of transcripts (coding and non-coding); (2) determine the transcriptional structure from the gene, such it is beginning and end sites (extremity 5' and 3', respectively), splicing patterns and others post-

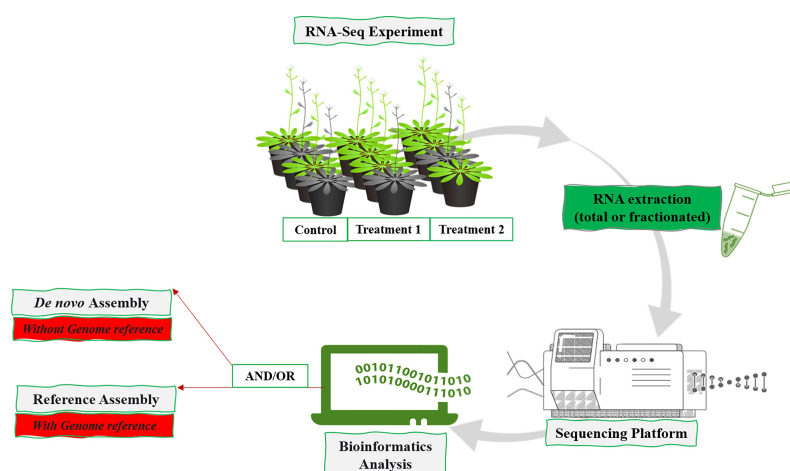
transcriptional modifications and; (3) quantify the expression level of each transcript in a given tissue, developmental stage and/or under different treatment conditions (Wang *et al.*, 2009).

In general, in this methodology, a sample of RNA (total or fractionated) is converted into a library of cDNA (complementary deoxyribonucleic acid) fragments with adapters connected to one (single-end sequencing) or both ends of the fragments (paired-end sequencing). Then, each molecule is then sequenced generating readings with fragments that vary in size depending on the platform used and the research objective (Wang *et al.*, 2009; Quail *et al.*, 2012).

The use of fractionated RNA could enrich a specific type of RNA, increasing the quantity of it by selection. Using this approach, it is possible to realize experiments based on coding RNAs by enriching mRNAs (messenger RNAs–protein encoders) or non-coding RNAs (gene expression regulators), such as miRNA (micro-RNA) (Chen *et al.*, 2019), mirtrons (one type of miRNA present in an intron) (Da Fonseca *et al.*, 2019) and lncRNA (Long non-coding RNA) (Negri *et al.*, 2019).

Currently, there are many sequencing platforms available. Among them, one example is illumina that is the most popular. This platform uses synthesis mechanism. Other example is the Chinese BGISEQ-500 from the BGI group; this platform applies technologies to sequence transcriptomes as cPAS (combinatorial Probe-Anchor Synthesis and DNA Nanoballs (DNB™) technology, showing competitiveness with Illumina platforms in quality and run price (AIOUB *et al.*, 2020).

Table 2 illustrates some of the sequencing platforms currently available and some technical specifications that may differ between them. In general, the choice of the sequencing platform should be based on the running cost and based on experimental design and research goals.












**Figure 1.** Overview of an RNA-Seq experiment and data analyses.

**Figura 1.** Visão geral de um experimento de RNA-Seq e análise dos dados.



**Table 2.** Sequencing platforms and their main features as mechanisms applied, size of reads, accuracy, running time, cost of equipment, and main analyses indication.

**Tabela 2.** Plataformas de sequenciamento e suas principais características como mecanismos, tamanho de fragmentos, acurácia, tempo de corrida, custo do equipamento e principal indicação.

Platform									
	<b>Sanger</b>	<b>HiSeq 2000</b>	<b>MiSeq</b>	<b>454</b>	<b>SOLiD v4</b>	<b>Ion Torrent</b>	<b>Genome Analyzer</b>	<b>PacBio</b>	<b>MinION</b>
Company	Applied Biosystem	Illumina	Illumina	Roche	Life Technologies	Life Technologies	Solexa	Pacific Bioscience	Oxford Nanopore
Generation	1st	2nd	2nd	2nd	2nd	2nd	2nd	3rd	4th
Mechanism	Synthesis	Synthesis	Syntheses	Pyrosequencing	Hybridization	Ions semi-conductors	Synthesis	Real-time synthesis	Nanopores
Reads (bp)	< 1 000	50-100	200	700	35-50	35-400	35-100	±15 000	±15 000
Accuracy (%)	99-99.9	99.9	99.9	99.9	99.9	99	98.5	>99%	99.8
Running time	3-15 days	3-10 days	2-3 days	1 day	7-14 days	4 h	5 days	0.5-4 h	2-3 h
Cost (thousand US\$)	N/A	690	128	500	495	80	430	695	1
Main application	Genome assembly	Gene expression; Splicing; SNP	Small genome assembly	Genome assembly; Epigenetics	SNP and Indels	Small genome assembly	Gene expression; Splicing	Genome assembly; Epigenetics; rare transcripts	Direct analyzes of RNA or cDNA
Author	Heather & Chain (2016)	Vincent <i>et al.</i> (2016)	Quail <i>et al.</i> (2012)	Vincent <i>et al.</i> (2016)	Vincent <i>et al.</i> (2016)	Quail <i>et al.</i> (2012)	Carvalho & Silva (2010)	Quail <i>et al.</i> (2012)	Jam <i>et al.</i> (2018)

## 2.3 Experimental design

The experimental design for sequencing depends on the type of analysis (splicing, SNP, and/or differential expression) to be performed and consists of defining the type of library (paired-end or single-end), the size of the fragments (reads) that will be obtained, the number of biological replicates, as well as the depth and coverage of the sequencing necessary to meet the proposed objectives based on the reference genome, if available (Conesa *et al.*, 2016).

For the analyzes of differentially expressed genes (DEGs), single-end libraries are sufficient to obtain the desired responses, however, for the analyzes of

alternative splicing (different forms of exon's junction) and analysis of SNP's (Single-nucleotide polymorphisms) it is essential to develop libraries with longer fragments and of the paired-end type (Volker & Small, 2017). The sequencing coverage should also be considered, since a stringent treatment in a data set with low coverage can result in a reduced number of fragments, making unfeasible and damaging later stages of analyzes and especially the identification of rare transcripts. Mainly whether the sequencing platform is based on PCR as Illumina. More detailed information is described in Table 3 (Volker & Small, 2017).

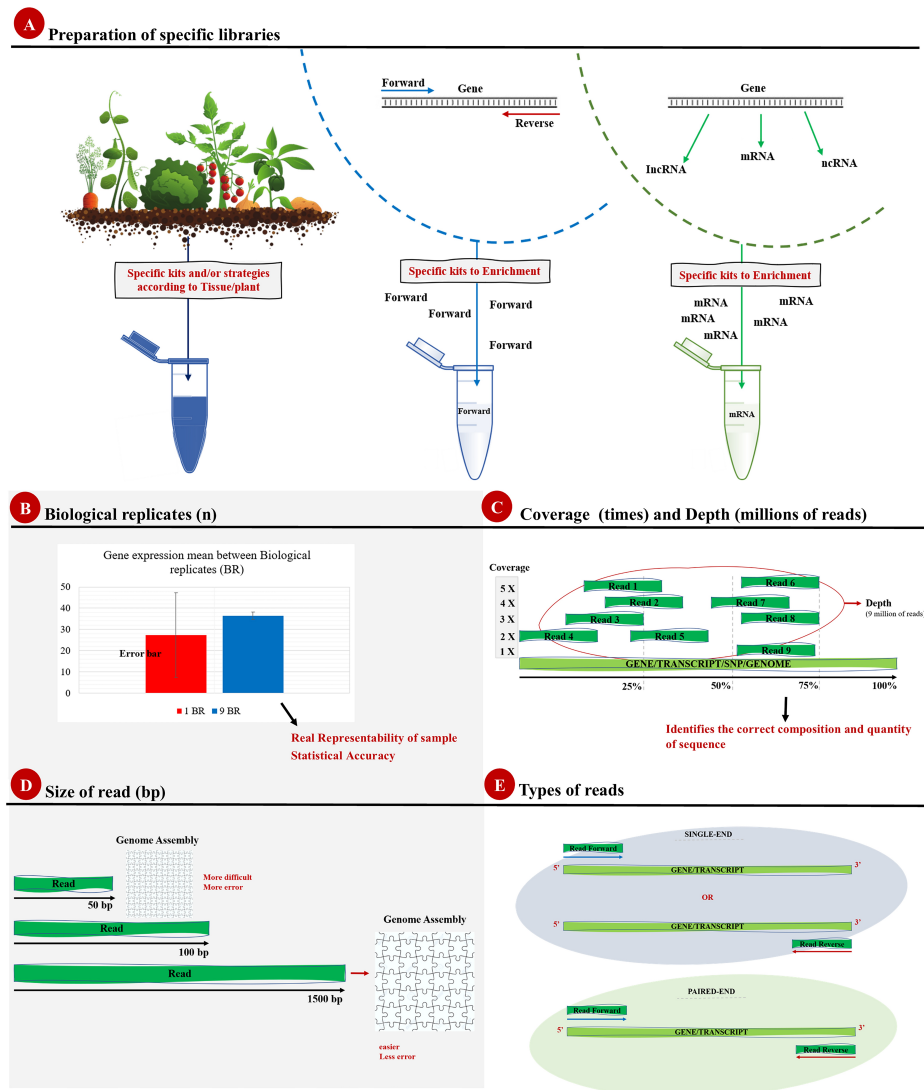
Figure 2 compiles all characteristics mentioned in Table 3 in an illustrated graphic to help to understand.

**Table 3.** Recommendations for RNA-Seq data analysis based on experimental objectives (Feng *et al.*, 2012; Conesa *et al.*, 2016; Volker & Small, 2017, adapted by the authors).

**Tabela 3.** Recomendações para análises de dados de RNA-Seq baseados nos objetivos experimentais (Feng *et al.*, 2012; Conesa *et al.*, 2016; Volker & Small, 2017, adaptado pelos autores).

Parameter	Splicing, <i>de novo</i> assembly, and SNP's analyzes	Differentially expression analyzes
Biological replicates (n)	Essential to have accurate statistics and representability of biological samples	As higher as better
Coverage (times)	Important. As higher as better, as more precise will be the genes/transcripts/SNP's quantification. By increasing coverage, errors from sequence base identification are reduced	As higher as better
Depth (millions of reads)	At least 1X genome coverage Calculate according to genome size and considering about 50% of reads loss during the cleaning procedure	Sufficient to obtain statistical accuracy and to decrease incorrect identification of bases
Size of reads (bp)	Important for <i>de novo</i> assembly (fragments >100 bp) and variants/isoforms identification (fragments from 100-250 bp). Long reads decrease assembly error of the contig. Small reads increase ambiguity (fragments <50 bp)	Normally not required, especially if a reference genome exists. Once exons splicing junctions are known (50-150 bp)
Reads paired-end	Important for <i>de novo</i> assembly and variants/isoforms identification as information comes from both sense strands	Normally not required, especially if a reference genome exists. The use of single-end reads is sufficient.
Preparation of specific libraries	Important for <i>de novo</i> assembly and identification of antisense transcripts. This approach provides enrichment of a certain type of RNA of interest (example mRNA) or even facilitate the identification of the sense that the gene is in the genome. Also, libraries can be tissue-specific.	Normally strand-specific not required, especially if a reference genome exists





**Figure 2.** Overview of RNA-Seq libraries features to be observed and defined according to the research objective. In A, preparation of specific libraries; in B, Biological replicates; in C, Coverage, and depth of an RNA-Seq; in D, size of the reads and in E, types of reads.

**Figura 2.** Visão geral das características das bibliotecas de RNA-Seq importantes a serem observadas e definidas de acordo com o objetivo da pesquisa. Em A, preparação de bibliotecas específicas; em B, réplicas biológicas; em C, cobertura e profundidade de um RNA-Seq; em D, tamanho das leituras e em E, tipos de leituras.

## 2.4 Transcriptome assembly with genome reference and *de novo* assembly

There is no standard pipeline established for the analyzes of this data, that is why it is always necessary to correctly plan the experimental design. If the design has already been carried out in the past, the choice of the appropriate software for the analysis is fundamental (Conesa *et al.*, 2016; Ezponda *et al.*, 2020).

To start the analysis, the first topic that must be considered is the availability of the genome or transcriptome of the organism under study. Check its availability allows the reference-based mapping, this approach reduces errors and computational demand, requiring less RAM (random access memory) memory using lighter algorithms if compared to the *de novo* assembly analyzes. *De novo* assembly is another approach that consists of assembling the genome and writing it down for the first time (Volker & Small, 2017).

Besides, the *de novo* assembly process is much more

laborious compared to mapping based on the reference genome, requiring greater coverage and depth since the programs used for this type of assembly make many more mistakes. One of the most used software-package used for *de novo* assembly is Velvet associated with Oases software (<https://github.com/dzerbino/velvet>) and Trinity (<https://omictools.com/trinity-tool>). This software establishes hash tables containing all sub-sequences of reads of different sizes (k-mer). In this step, the size of the k-mer and the type of reading must be informed for the construction of Bruijn graphs, which allow the visualization of overlapping k-mers and permit the construction of contigs from these overlaps (Conesa *et al.*, 2016).

Although the *de novo* assembly requires a greater computational capacity and presents a relatively high error rate, it is the only methodology that enables the identification of new transcripts, since the genome reference-based assembly does not cover such a possibility, it only identifies the expression of transcripts

already known/previously noted (Conesa *et al.*, 2016).

## 2.5 Data input and treatment of raw data - Trimming/cleaning and sequences quality control

The first step in transcriptome analysis, using RNA-Seq data, is the visualization of raw data (output directly from sequencers) to check the quality of the obtained sequences and the identification of possible problems that could result from both the biological samples extraction and from the sequencing-run. FastQC is a bioinformatic tool widely used for this purpose. In short, this software extracts all the important information mentioned above and issues a report in HTML format containing all the main quality graphics generated, in a free and friendly graphic interface, allowing data viewing, but not the data edition. As a standard procedure FastQC software is applied before and after the cleaning of the reads to check sequence quality (Conesa *et al.*, 2016).

With the report issued by FastQC, the next step is to clear the reads, which efficiently reduces alignment and contigs assembly error rates. This process comprises three procedures: the removal of adapters, the elimination of the sequences with low quality, and the recovery of sequences with desirable length. Adapters are sequences of known bases, present at both (paired-end), and only one (single-end) end. The complete removal of adapters is important because their sequences can be incorporated in the alignment of reads to the genome, resulting in an erroneous assembly of the sequences reads (Wang *et al.*, 2009; Quail *et al.*, 2012).

Most reports in the literature have established as a standard to remove low-quality sequences, Phred scores between 20 and 30, which represent the accuracy of 99 and 99.9%, respectively (Prosdociimi, 2006). Finally, it is recommended the discard short sequences (less than 40 bp), once they occur several times within the target sequence and, therefore, may result in ambiguous information (Bolder *et al.*, 2014). Several tools can be applied to trim/clean the sequences such as PRINSEQ (<http://prinseq.sourceforge.net/>), FastX-Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)), TagCleaner (<http://tagcleaner.sourceforge.net/>), Cutadapt (<https://cutadapt.readthedocs.io/en/stable/>), Trimmomatic (Bolder *et al.*, 2014) and SeqClean (<https://github.com/ibest/seqclean>), depending on the sequencing platform used.

Trimmomatic, for example, is a software developed for cleaning data from the Illumina platform which avoids excessive tabulation overhead and clears raw data very quickly. It is a flexible tool for reliable data management for both single-end and paired-end libraries, for reference genome assembly or *de novo* assembly (Bolger *et al.*, 2014). This software includes a variety of processing steps, but the main algorithmic innovations are related to the identification of adapters and quality filtering (Bolger *et al.*, 2014). To perform it, the software uses the Illumina platform quality score for each base position to determine where the reading should be cut, resulting in the retention of the 5' portion, while the 3' cut off sequence is

discarded. This analysis is quite convenient for typical Illumina platform data as these generally display lower quality at 3' end extremity. This information makes it clear that software with algorithms better suited to certain sequencing platforms should be used.

## 2.6 Differential gene expression analyzes and alternative splicing

In general, pipelines for the analyzes of differentially expressed genes or alternative splicing, based on alignment with the reference genome, use input data containing millions of reads from the transcriptome (FASTQ file), from the reference genome (FASTA file), and genome annotation (GFF3 or GTF file - gene transfer format). These generate output data is usually resulting from the alignment (SAM - sequence alignment map and BAM - binary alignment map), assembly (GFF3 - general feature format 3 or GTF), from tables containing the differentially normalized genes (Log2FC - logarithm 2 fold change) and/or containing the alternative splicing rate (PSI file).

Following sequences quality check and trimming, the next step is to perform the alignment. To carry out the alignment with the reference genome, some of the software used are MAQ (<http://maq.sourceforge.net/maq-man.shtml>), BWA (<http://bio-bwa.sourceforge.net/>), Bowtie2 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>), TopHat2 (<https://ccb.jhu.edu/software/tophat/index.shtml>), STAR (<https://github.com/alexdobin/STAR>), HISAT2 (Kim *et al.*, 2015) among others. These software use strategies such as BWT-burrows-wheeler dispersion and transformation tables that have the following statements as computational premises: the fragments are small, there are millions of them, and there are sequencing errors and repetitive regions in the genomes. All these characteristics make assembly difficult and require the development of efficient computational algorithms and tools. In addition to alignment, BWT is a reversible sequence permutation technique, initially developed for data compression, that can create indexes of a genome using little computational space. These indices, such as FM-index, can be used to search for sequences in a reduced domain of sub-sequences, without the need to search in the entire genome (Kim *et al.*, 2015).

According to Kim *et al.* (2015), the HISAT software is the fastest system currently available for sequence alignment based on a reference genome, with better precision and more quickly compared to its predecessor, the TopHat2 software. HISAT requires only 4.3 gigabytes (GB) of RAM and is capable of aligning genomes of any size, including those larger than 4 billion bases, such as the sunflower and corn genomes, and the human genome, among others. Another widely used software is STAR (Dobin *et al.*, 2013), which unlike others, uses suffix arrays to perform data processing, however, this requires very large memory; for the human genome, for example, it required 28 GB of RAM, slower than other methods that apply the Burrows-Wheeler transformation, such as

HISAT2 (Kim *et al.*, 2015). In this context, the HISAT2 software can be used by users who do not have access to servers and need to perform their analyzes with quality, using little RAM on their personal computers.

The HISAT2 software is a Gapped mapper aligner, that is, it can identify exon junctions, allowing to evaluate, in addition to the differential expression, the occurrence of alternative splicing (Kim *et al.*, 2015). Splicing is a process that involves the removal of introns and the union of exons after RNA transcription, and combinations of different forms (alternative splicing) can occur, thus resulting in the encoding of multiple proteins from the same gene (Li *et al.*, 2017). Although HISAT2 software has this applicability, there is software designed exclusively for this type of analyzes such as SUPPA2 (<https://github.com/comprna/SUPPA>). Other software available for alternative splicing analyzes is ASTALAVISTA (<http://astalavista.sammeth.net/>) and ASpli (<https://bioconductor.org/packages/release/bioc/html/ASpli.html>). The choice of the ideal software for this analyzes depends on the capacity of the software, the computational capacity required to run the data, and the type of parameter applied to identify the different types of splicing (Carazo *et al.*, 2018). A review work carried out by Carazo *et al.* (2018), described 33 specific software for alternative splicing analyzes, which works in different operating systems, require different computational capacities, and are capable of identifying or not new variants, among other parameters.

After reads alignment in the reference genome, it is possible to assembly the transcripts. To this end, tools such as Cufflinks (<http://cole-trapnell-lab.github.io/cufflinks/>) and Stringtie (<https://ccb.jhu.edu/software/stringtie/>) are widely used. Literature data show that the StringTie program allows the correct assembly of a larger number of transcripts twice as efficiently and quickly when compared to Cufflinks and other programs. Briefly, the assembly process consists of overlapping the reads to form continuous stretches of DNA, called contigs, which will generate larger fragments that will be assembled in supercontigs (scaffolds) and thus finally, in transcripts (Conesa *et al.*, 2016).

Frequently, during all stages of RNA-Seq data analysis, the use of a specific software within a pipeline may require converting files into different formats once each software requires a specific input file type. SAMTools (<http://samtools.sourceforge.net/>) is a toolkit that converts files into different formats. Also, this package presents a tool that removes PCR artifacts (true unnatural duplicates) that can also cause mapping errors. Unnatural duplicates arise from multiple cycles of PCR that can generate molecules with patterns that do not reflect the information of the template molecule and their removal aims to reduce noise and minimize the identification of false-positive genes. The most used packages in the literature to identify and remove these duplicates are the MarkDuplicates tools, from the Picard package (<http://picard.sourceforge.net/>) and rmdup from the

SAMTools package. The difference between these packages is in the way that the algorithms are applied to the data; MarkDuplicates does not exclude duplicates if not requested, allowing only their visualization, so the user can choose to consider the duplicates or not in the following analyzes, requiring however more RAM to run; rmdup instead removes duplicates quickly and uses little RAM.

## 2.7 Normalization algorithms to read counts comparison within and among libraries

Not only is the cleaning and removal of PCR artifacts important to generate highly reliable data, but also the selection of normalization algorithms that must be compatible with the type of library (single or paired-end) and the number of biological repetitions and treatments. Several algorithms can be used, each with its advantages, disadvantages, and the most appropriate use. In any way, the outcoming data from the sequencing needs to be normalized to remove the biases inherent to the technique, mainly regarding the length of the genes and the sample sequencing depth. These biases are mostly corrected within a sample by RPKM (reads per kilobase per million reads mapped) in single-end libraries and FPKM (fragments per kilobase per million reads mapped) for paired-end libraries. Another index used, TPM (transcripts per million), considers in this correction the length of the genes and the library sequencing depth, allowing the visualization of the real expression levels of one gene about another and/or between different libraries, since this algorithm considers in the analyzes the different depths between libraries (Wagner *et al.*, 2012). It is important to highlight that the RPKM does not respect the invariance property and, therefore, cannot be an accurate measure of relativity. The TPM is a modification of the RPKM that eliminates this inconsistency, respecting the average invariance, and eliminating the statistical prejudices inherent to the RPKM index (Wagner *et al.*, 2012). This type of normalization has already been widely used in several areas of knowledge (Machado *et al.*, 2020). The RPKM formula is  $RPKM = C / LN$  where: • C: Number of fragments mapped to a sequence (i.e., transcript, exon, etc.). • L: Length of the sequence (in kb). • N: Total number of mapped fragments (in millions) (Wagner *et al.*, 2012). The TPM formula is:  $TPM = (N \times 1000000) / (T \times L)$  where: • N: Number of fragments mapped to a sequence (i.e., transcript, exon, etc.). • L: Length of the sequence (in Kb) • T: Sum of all transcription fees (Sum of N / L).

It is also important to emphasize that to have this equal proportion, the TPM uses a T (Scaling factor) value, which represents the sum of all normalized transcription expression values divided by 1,000,000. Another normalizer based on counts, CPM (counts per million), contemplate differences in the size of libraries, however, these counts are calculated as gross counts divided by the sizes of the libraries and multiplied by one million, so the proportionality of comparison of the gene expression can be preserved (Robinson *et al.*, 2010). There are also other



widely used normalizations, such as the normalization in TMM (trimmed mean of m-values) described and implemented in the edgeR package; the normalization by the expression of RLE (relative log expression) applied in the DESeq2 package and finally MRN (normalization median ratio normalization) (Maza *et al.*, 2016).

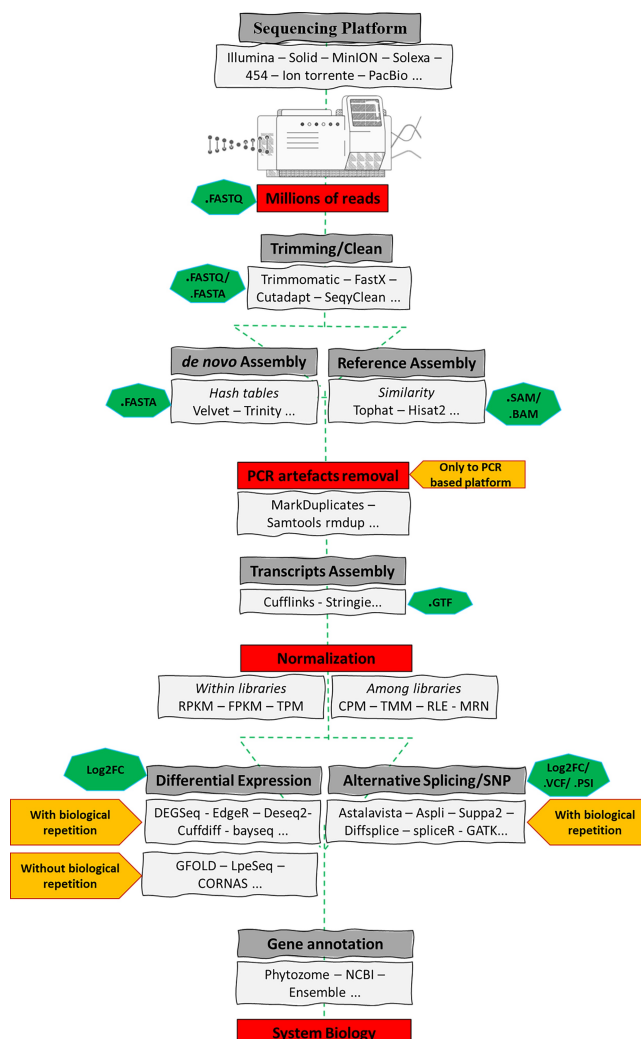
Normalizations are usually performed in the stages of mapping and analysis of differential expression; in this last stage, the differences in the level of gene expression between the different treatments are observed on a scale in Log2FC (logarithm 2 of fold change), which is a value calculated in the logarithm of base 2 (log2FC). For example, log2FC with values 1/-1 represents an expression twice as higher/lower for a given sample about another. To this end, statistical tests are applied to the summarized read count data for each library. As count data are discrete variables, some discrete probability distributions are more suitable for obtaining information from RNA-Seq data, such as the Poisson distribution and the negative binomial distribution (Conesa *et al.*, 2016). Most of the available software to perform differential expression analyzes require three biological repetitions to generate accurate data and consequently robust results.

However, the limited availability of samples and/or financial resources generally results in studies with small sample sizes, or even, with only one biological repetition.

Nevertheless, with the evolution of bioinformatics tools, it became possible to obtain differential gene expression even from a limited number of samples, in particular samples without replicates. To solve this bias, there is specific free software such as GFOLD (Feng *et al.*, 2012) and LPESeq

(<https://github.com/JungsooGIM/LPEseq/wiki>). These two software were compared with others already well established for RNA-Seq analyzes, such as DESeq, DEGSeq, and edgeR, and were shown to be more efficient when there is only one biological replicate (Feng *et al.*, 2012). In particular, the GFOLD software has been widely used in several areas of knowledge, showing a high correlation with results from RT-qPCR validation (Oh *et al.*, 2020).

In summary, the pipeline for the assembly of transcripts using a reference genome consists of the cDNA libraries cleaning step, followed by the alignment, and the mapping of the differential expressed genes and/or alternative splicing, as shown in Figure 3.



**Figure 3.** Overview of the differential gene expression analyzes pipeline and alternative splicing in RNA-Seq experiments.

**Figura 3.** Visão geral do pipeline de análise de expressão diferencial e *splicing* alternativo em experimentos de RNA-Seq.

## 2.8 Systems Biology

After a list containing the genes differentially expressed in a given condition or tissue is obtained, systems biology approaches can be applied (Oshlack *et al.*, 2010). The first step after identifying the differentially expressed genes is their functional annotation. For this purpose, several global databases are available, such as NCBI's Genbank (<https://www.ncbi.nlm.nih.gov/>), specific databases for plants like Phytozome (<https://phytozome.jgi.doe.gov/pz/portal.html>) and databases exclusive for certain species such as Arabidopsis (TAIR - <https://www.arabidopsis.org/>), soybean (Soybase - [https://www.soybase.org/goslimgraphic\\_v2/dashboard.php](https://www.soybase.org/goslimgraphic_v2/dashboard.php) and SoyKB - <http://soykb.org/>), maize (<https://www.maizegdb.org/>), rice (<http://iric.irri.org/resources/rice-databases/>), wheat and oat (<https://wheat.pw.usda.gov/GG3/>) among others.

Also, specific types of RNAs have different databases and software used to annotate and analyze data. As examples, we can cite mirtronDB, the first database dedicated to mirtron, available at <http://mirtrondb.cp.utfpr.edu.br/>. This database currently contains a total of 1.407 mirtron precursors and 2.426 mirtron mature sequences identified in 18 different species. To analyze different types of miRNA, diverse different software are available, and the comparison between them is available at <https://doi.org/10.1093/bib/bby054>. For long non-coding RNA (lncRNAs), a machine learning analysis and a classification approach called RNAplonc is available at <https://github.com/TatianneNegri/RNAplonc/>. Usually, databases are classified based on the information source, the type of RNA, organisms, the data formats, and the mechanisms for information retrieval. This information is detailed in a book chapter described by Maracaja-Coutinho *et al.* (2019), in which the relevance of each of all these classifications and its use by researchers is addressed. Besides that, many software and more information about the mechanism of gene regulation by non-coding RNA activity are available at <https://doi.org/10.1093/bib/bbx058>.

After biological annotation, other analyzes such as ontological categorization can be carried out on sites such as Gene Ontology Consortium (<http://geneontology.org/>), CateGORizer (<https://www.animalgenome.org/tools/catego/>), AgriGO (<http://amigo.geneontology.org/amigo>) and MapMan (Thimm *et al.*, 2004). Unfortunately, many GO-Gene Ontology enrichment analyzes do not consider the expression of the genes, which is not interesting when it is desired to obtain the maximum representativeness of the worked data. AgriGO software focuses on agricultural species, it is free, extremely friendly (Tian *et al.*, 2017), and it features a tool called PAGE-Parametric Analyzes of Gene set Enrichment, which consists of cross-comparisons of multiple data sets and GO annotations of the reference genome, contemplating in the analyzes not only the number of genes in a given category but also the expression values

(<http://bioinfo.cau.edu.cn/agriGO/analizes.php?method=PAGE>). To calculate the significance of the expression values average from genes grouped into the same GO, this software applies the Hochberg FDR test as a statistical method with a level of significance based on adjusted p-value (Tian *et al.*, 2017).

Mapman is another software option that categorizes the input data using updated versions of the genome of interest or newly *de novo* assembly for annotation, also performing expression analyzes from the log2FC values, in a simple Java graphical interface, which generates illustrated figures of the main metabolic pathways observed from the input data (Thimm *et al.*, 2004). Although this software has an excellent proposal, a limiting factor in the analyzes are libraries displaying few genes, since the smaller the number of genes, the less likely they are to have a GO assigned. The default parameters for some of the available software required at least 5 GOs terms for a gene to be represented in the generated images (Thimm *et al.*, 2004).

Besides, exploratory, and visual analyzes, such as the creation of heatmaps, can be performed by tools like Morpheus (<https://software.broadinstitute.org/morpheus/>), Heatmapper (<http://www.heatmapper.ca/>) and ClustVis (<https://biit.cs.ut.ee/clustvis/>). The latter also enables the analyzes of PCA (principal component analyzes) based on the R language. The PCA is a way of identifying the relationship between characteristics gotten from the data, showing the arrangement that best represents the data distribution.

There are also tools for rescue and analysis of promoter sequences such as RSAT (<http://rsat.eead.csic.es/plants/>), Promoter Inspector, Promoter Scan ([https://bip.weizmann.ac.il/education/ws/0203/211102\\_sbd/promoterworkshop.html](https://bip.weizmann.ac.il/education/ws/0203/211102_sbd/promoterworkshop.html)), and PlantPAN 2.0 (<http://plantpan2.itps.ncku.edu.tw/promoter.php>). With these sequences it is possible to carry out analyzes of known cis-elements present in this gene region, using software such as PlantCare (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>) and PLACE (<http://www.dna.affrc.go.jp/PLACE/>) or discover new cis-elements with the software MEME suits (<http://meme-suite.org/tools/meme>).

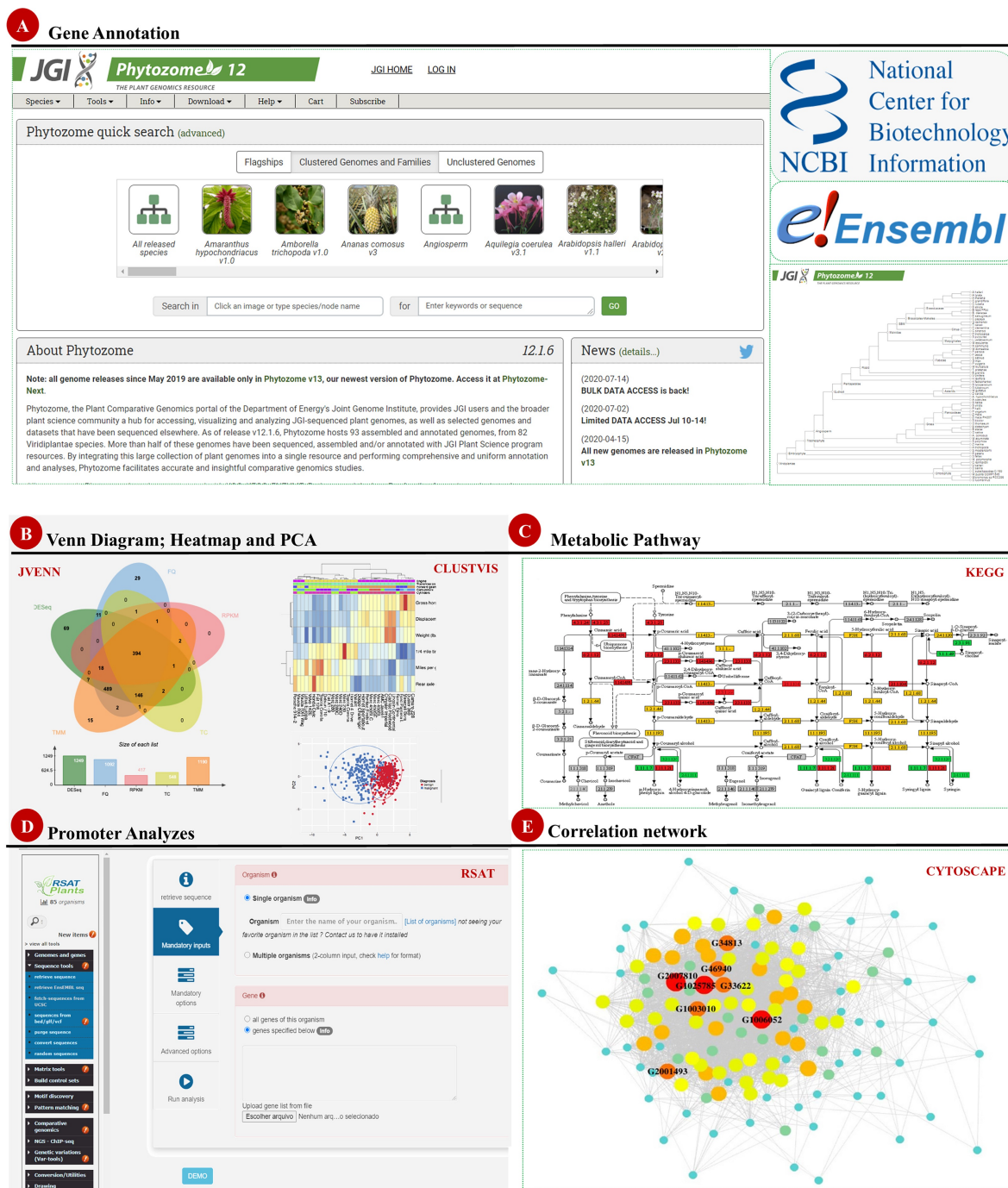
Broader analyzes of the total set of differentially expressed genes can also be carried out by identifying metabolic networks, using databases such as the KEGG pathway (<https://www.genome.jp/kegg/>), Kbase (<http://kbase.us/metabolic-modeling-in-kbase/>), and Pathway tools (<http://bioinformatics.ai.sri.com/ptools/>). Also, analyzes of gene correlation networks can be made, using software such as Cytoscape (<https://cytoscape.org/>), String (<https://string-db.org/>), and SoyCSN (<http://soykb.org/SoyCSN/>).

Genes differentially expressed exclusively and shared among the analyzed samples can be identified and visualized in Venn diagrams using software such as Venny (<https://bioinfogp.cnb.csic.es/tools/venny/>), Venn diagram (<http://bioinformatics.psb.ugent.be/webtools/Venn/>), Jvenn (<http://jvenn.toulouse.inra.fr/app/index.html>), and InteractiVenn (<http://www.interactivenn.net/>). Still, from the gene sequences, it is also possible to perform protein

analyzes to identify whether the proteins encoded by the genes are active and present or not conserved domains. It is likewise possible to visualize the 3D structure of these proteins through software such as Castp (<http://sts.bioe.uic.edu/castp/index.html?2cpk>) and a

database such as PDB (protein data bank) (<https://www.wwpdb.org/>).

Some possibilities of biological systems analyzes are shown in Figure 4.



**Figure 4.** Some possible systems biology analyzes that can be performed using RNA-Seq data. In A, Phytozome, Ensembl and NCBI websites, used to carry out gene annotation, among other analyses possibilities. In B, different ways to present data such as Venn Diagram (analyses carry out at JVENN software), HeatMap, and PCA (principal component analyses); performed at CLUSTVIS. In C, an example of a non-dynamic metabolic pathway from the KEGG website. Image provided by Alana Madureira. In D, promoter sequence retrieval and analysis were performed at RSAT. In E, an analysis example of correlation network was carried out at CYTOSCAPE software. Image from <http://dx.doi.org/10.1101/496075>.

**Figura 4.** Possibilidades de análises de biologia de sistemas que podem ser realizadas a partir de dados de RNA-Seq. Em A, os sites Phytozome, Ensembl e NCBI, utilizados para realizar anotações de genes, entre outras possibilidades de análises. Em B, diferentes formas de apresentação dos dados como Diagrama de Venn (análises realizadas no software JVENN), HeatMap e PCA (análises de componentes principais) realizadas no CLUSTVIS. Em C, exemplo de via metabólica não dinâmica do site KEGG. Imagem fornecida por Alana Madureira. Em D, recuperação e análise da sequência do promotor realizada no RSAT. Em E, um exemplo de análise de rede de correlação realizada no software CYTOSCAPE. Imagem de <http://dx.doi.org/10.1101/496075>.



Currently, there is a large amount of public RNA-Seq database available in the literature for several species, and in addition to the raw data, information from the experimental design can be retrieved in an organized approach in databases such as NCBI's SRA (sequence read file) (<https://www.ncbi.nlm.nih.gov/sra>). As an example, specifically for soybeans, it is also possible to retrieve data already standardized in TPM from 1298 RNA-Seqs of different tissues in the Soybean Expression Atlas bank ([http://200.20.229.99/cgi-bin/gmax\\_atlas/index.cgi](http://200.20.229.99/cgi-bin/gmax_atlas/index.cgi)). In this context, bioinformatics analyzes can be performed, not only based on data from experiments designed by the author himself, but also by rescuing previously sequenced libraries from these databases (Conesa *et al.*, 2016; Min *et al.*, 2020).

The purpose of all these computational analyzes, which through algorithms designed specifically for each type of analysis (differential expression; alternative splicing; SNP's) and experimental design (a type of library; biological repetition; coverage; depth; read size; reference genome) are to increase the assertiveness of true biological responses and consequently generate savings on bench reagents during validation.

### 3 Final Considerations

Bioinformatics emerged as an interdisciplinary field to help analyze the enormous data generated by the current sequencing technologies. Although bioinformatics tools provide us a huge amount of software options to apply in the analyses, some important question regarding experimental design, biological replicates, and research objectives among others, should be addressed to perform a robust and reliable analysis and obtain results that represent the true biological responses involved in the conditions of interest.

In this review, a step by step analysis pipeline was described for RNA-Seq experiments with and without biological replicates. By considering important and decision questions, in each stage of the analyses, the authors hope to clarify and provide the scientific community, detailed search material to develop trustworthy research and reach consistent results, with a high correlation with other gene expression quantification techniques such as RT-qPCR.

### References

- AIIOUB, A.A.; ZUO, Y.; LI, Y.; QIE, X.; ZHANG, X.; ESSMAT, N.; HU, Z. Transcriptome analysis of *Plantago* major as a phytoremediator to identify some genes related to cypermethrin detoxification. **Environmental Science and Pollution Research**, v. 27, p. 1-15, 2020. <https://doi.org/10.1007/s11356-020-10774-4>
- BOLGER, A.M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014. <https://doi.org/10.1093/bioinformatics/btu170>
- CAMARENA, L.; BRUNO, V.; EUSKIRCHEN, G.; POGGIO, S.; SNYDER, M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. **PLoS pathogens**, v. 6, n. 4, p. e1000834, 2010. <https://doi.org/10.1371/journal.ppat.1000834>
- CARAZO, F.; ROMERO, J.P.; RUBIO, A. Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. **Briefings in Bioinformatics**, v. 20, n. 4, p. 1358-1375, 2019. <https://doi.org/10.1093/bib/bby005>
- CHEN, L.; HEIKKINEN, L.; WANG, C.; YANG, Y.; SUN, H.; WONG, G. Trends in the development of miRNA bioinformatics tools. **Briefings in Bioinformatics**, v. 20, n.5, p. 1836-1852, 2019. <https://doi.org/10.1093/bib/bby054>
- CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; MORTAZAVI, A. A survey of best practices for RNA-Seq data analyzes. **Genome biology**, v. 17, n. 1, p. 13, 2016. <https://doi.org/10.1186/s13059-016-0881-8>
- DA FONSECA, B.H.R.; DOMINGUES, D.S.; PASCHOAL, A.R. mirtronDB: a mirtron knowledge base. **Bioinformatics**, v.35, n. 19, p. 3873-3874, 2019. <https://doi.org/10.1093/bioinformatics/btz153>
- EVERAERT, C.; LUYPAERT, M.; MAAG, J. L.; CHENG, Q. X.; DINGER, M. E.; HELLEMANS, J.; MESTDAGH, P. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. **Scientific Reports**, v. 7, n. 1, p. 1559, 2017. <https://doi.org/10.1038/s41598-017-01617-3>
- EWING B; GREEN, P. Base-Calling of automated sequencer traces using Phred. II. Error probabilities. **Genome Research**, v. 8, p. 186-194, 1998. <http://doi.org/10.1101/gr.8.3.186>
- FENG, J.; MEYER, C. A.; WANG, Q., LIU; J. S.; SHIRLEY LIU, X.; ZHANG, Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-Seq data. **Bioinformatics**, v. 28, n. 21, p. 2782-2788, 2012. <https://doi.org/10.1093/bioinformatics/bts515>
- HEATHER, J.M.; CHAIN, B. The sequence of sequencers: the history of sequencing DNA. **Genomics**, v. 107, n. 1, p. 1-8, 2016. <https://doi.org/10.1016/j.ygeno.2015.11.003>
- JAIN, M.; KOREN, S.; MIGA, K. H.; QUICK, J.; RAND, A. C.; SASANI, T. A.; MALLA, S. Nanopore sequencing and assembly of a human genome with ultra-long reads. **Nature Biotechnology**, v. 36, n. 4, p. 338-345, 2018. <https://doi.org/10.1038/nbt.4060>

- KIM, D.; LANGMEAD, B.; SALZBERG, S.L. HISAT: a fast-spliced aligner with low memory requirements. **Nature Methods**, v. 12, n. 4, p. 357-360, 2015. <https://doi.org/10.1038/nmeth.3317>
- LI, Q. Q.; LIU, Z.; LU, W.; LIU, M. Interplay between alternative splicing and alternative polyadenylation defines the expression outcome of the plant unique OXIDATIVE TOLERANT-6 gene. **Scientific Reports**, v. 7, n. 1, p. 1-9, 2017. <https://doi.org/10.1038/s41598-017-02215-z>
- MACHADO, F. B.; MOHARANA, K. C.; ALMEIDA-SILVA, F.; GAZARA, R. K.; PEDROSA-SILVA, F.; COELHO, F. S.; VENANCIO, T. M. Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. **The Plant Journal: for cell and Molecular Biology**, v. 103, n.5, p. 1894-1909, 2020. <https://doi.org/10.1111/tpj.14850>
- MARACAJA-COUTINHO, V.; PASCHOAL, A. R.; CARIS-MALDONADO, J. C.; BORGES, P. V.; FERREIRA, A. J.; DURHAM, A. M. Noncoding RNAs databases: current status and trends. In *Computational Biology of Non-Coding RNA*. Humana Press, New York, NY, p. 251-285, 2019. [https://doi.org/10.1007/978-1-4939-8982-9\\_10](https://doi.org/10.1007/978-1-4939-8982-9_10)
- MAZA, E. In Papyro comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-Seq experimental design. **Frontiers in genetics**, v. 7, p. 164, 2016. <https://doi.org/10.3389/fgene.2016.00164>
- MIN, J.; WAGNER, M.; KASAMIAS, T. Advances in Transcriptome Analyses Using RNA Sequencing Technology in Soybean Plants [*Glycine max*]. **Computational Molecular Biology**, v. 10, n. 1, 2020.
- MUHAMMAD, I. I.; KONG, S. L.; AKMAR ABDULLAH, S. N.; MUNUSAMY, U. RNA-seq, and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional Regulatory Mechanism. **International Journal of Molecular Sciences**, v. 21, n. 1, p. 167, 2020. <https://doi.org/10.3390/ijms21010167>
- NEGRI, T. D. C.; ALVES, W. A. L.; BUGATTI, P. H.; SAITO, P. T. M.; DOMINGUES, D. S.; PASCHOAL, A. R. Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants. **Briefings in Bioinformatics**, v. 20, n.2, p. 682-689, 2019. <https://doi.org/10.1093/bib/bby034>
- OH, J. M.; VENTERS, C. C.; DI, C.; PINTO, A. M.; WAN, L.; YOUNIS, I.; DREYFUSS, G. U1 snRNP regulates cancer cell migration and invasion in vitro. **Nature Communications**, v. 11, n. 1, p. 1-8, 2020. <https://doi.org/10.1038/s41467-019-13993-7>
- OSHLACK, A.; ROBINSON, M.D.; YOUNG, M.D. From RNA-Seq reads to differential expression results. **Genome Biology**, v. 11, n. 12, p. 220, 2010. <https://doi.org/10.1186/gb-2010-11-12-220>
- PROSDOCIMI, F.; DE CARVALHO, D. C.; DE ALMEIDA, R. N.; BEHEREGARAY, L. B. The complete mitochondrial genome of two recently derived species of the fish genus *Nannoperca* (Perciformes, Percichthyidae). **Molecular Biology Reports**, v. 39, n. 3, p. 2767-2772, 2012. <https://doi.org/10.1007/s11033-011-1034-5>
- QUAIL, M. A.; SMITH, M.; COUPLAND, P.; OTTO, T. D.; HARRIS, S. R.; CONNOR, T. R.; GU, Y. A tale of three next-generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences, and Illumina MiSeq sequencers. **BMC Genomics**, v. 13, n. 1, p. 341, 2012. <https://doi.org/10.1186/1471-2164-13-341>
- SIMPSON, A. J. G.; REINACH, F. D. C.; ARRUDA, P.; ABREU, F. A. D.; ACENCIO, M.; ALVARENGA, R.; BARROS, M. H. D. The genome sequence of the plant pathogen *Xylella fastidiosa*. **Nature**, v. 406, n. 6792, p. 151-157, 2000. <https://doi.org/10.1038/35018003>
- THIMM, O.; BLÄSING, O.; GIBON, Y.; NAGEL, A.; MEYER, S.; KRÜGER, P.; STITT, M. Mapman: a user# driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. **The Plant Journal**, v. 37, n. 6, p. 914-939, 2004. <https://doi.org/10.1111/j.1365-313X.2004.02016.x>
- TIAN, T.; LIU, Y.; YAN, H.; YOU, Q.; YI, X.; DU, Z.; SU, Z. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. **Nucleic acids research**, v. 45, n. 1, p. 122-129, 2017. <https://doi.org/10.1093/nar/gkx382>
- VOLKER, R.; SMALL, C. *RNA-seqlopedia*, 2017. Disponível em: <https://RNA-Seq.uoregon.edu/#exp-design>. Acesso em: 09 out. 2019.
- WAGNER, G.P.; KIN, K.; LYNCH, V.J. Measurement of mRNA abundance using RNA-Seq data: RPKM measure is inconsistent among samples. **Theory in biosciences**, v. 131, n. 4, p. 281-285, 2012. <https://doi.org/10.1007/s12064-012-0162-3>
- WANG, M.; JIANG, B.; LIU, W.; LIN, Y. E.; LIANG, Z.; HE, X.; PENG, Q. Transcriptome Analyzes Provide Novel Insights into Heat Stress Responses in Chieh-Qua (*Benincasa hispida* Cogn. var. Chieh-Qua How). **International journal of molecular sciences**, v. 20, n. 4, p. 883, 2019. <https://doi.org/10.3390/ijms20040883>
- WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57-63, 2009. <https://doi.org/10.1038/nrg2484>

ZHANG, C.; DOWER, K.; ZHANG, B., MARTINEZ; R. V., LIN; L. L.; ZHAO, S. Computational identification, and validation of alternative splicing in ZSF1 rat RNA-seq data, a preclinical model for type 2 diabetic nephropathy. **Scientific Reports**, v. 8, n. 1, p. 7624, 2018. <https://doi.org/10.1038/s41598-018-26035-x>

**Acknowledgements:** The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPQ) and Araucaria Foundation for granting scholarships to some of the authors.

**Author's contribution statement:** Mayla Daiane Correa Molinari: Conceptualization, Writing -original draft, Visualization - First Writing; Renata Fuganti-Pagliarini: Conceptualization, Writing - review & editing, Visualization – First Writing. Jéssika Angelotti-Mendonça: Writing - review & editing. Daniel de Amorim Barbosa: Writing - review & editing; Daniel Rockenbach Marin: Writing - review & editing. Liliane Mertz-Henning: Supervision, Writing - review & editing; Alexandre Lima Nepomuceno: Supervision, Writing - review & editing.

**Conflict of Interest:** The authors declare no conflict of interest.

**Section Editor:** Lucas da Silva Santos